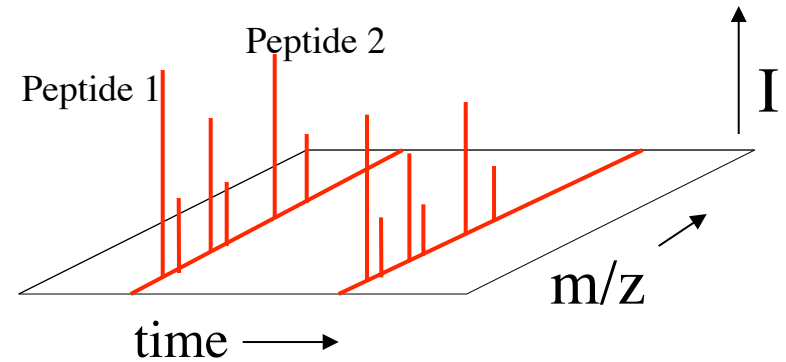
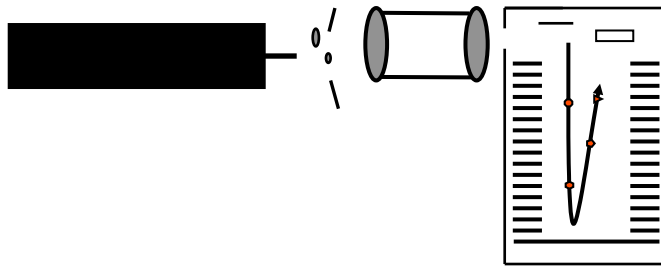


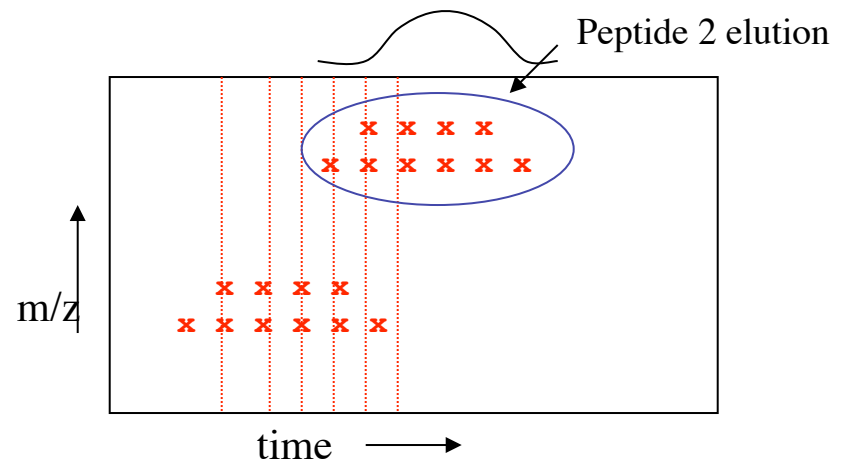
L14

Mass Spec Quantitation  
MS applications  
Microarray analysis

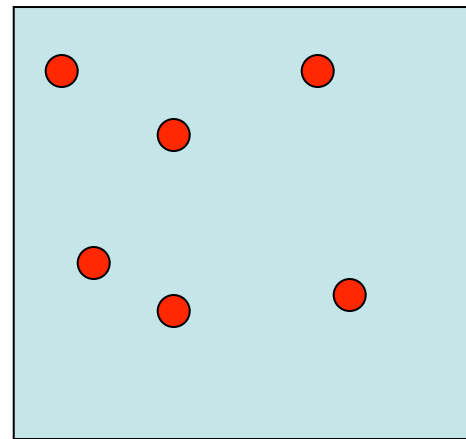
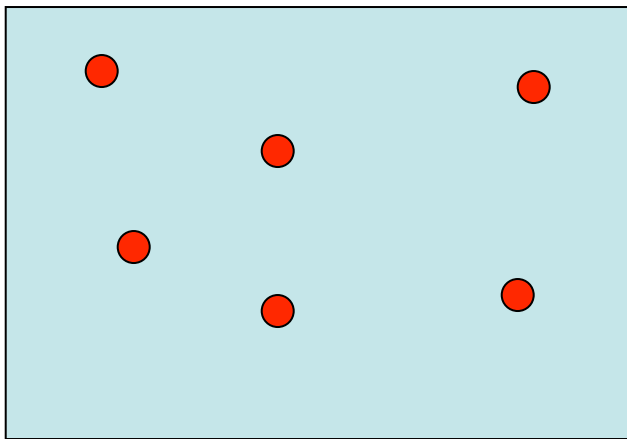
# LC-MS Maps



- A peptide/feature can be labeled with the triple (M,T,I):
  - monoisotopic M/Z, centroid retention time, and intensity
- An LC-MS map is a collection of features



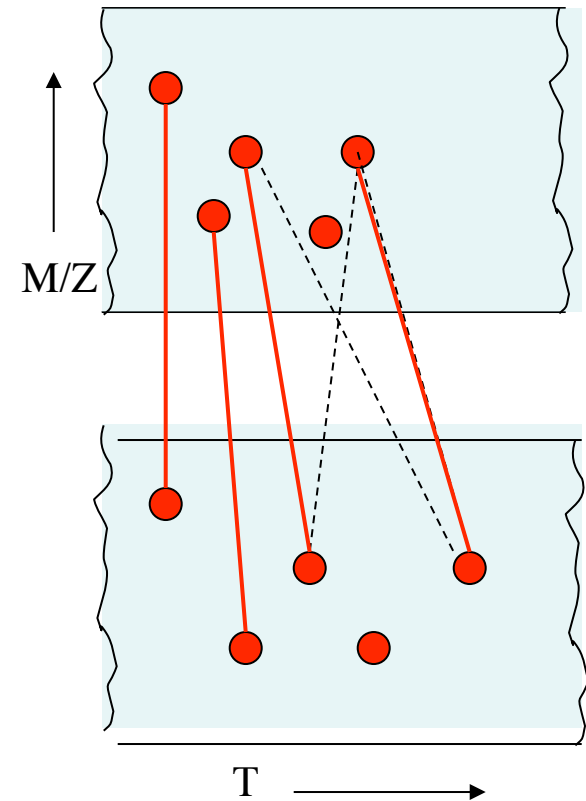
# Time scaling: Approach 1 (geometric matching)



- Match features based on  $M/Z$ , and (loose) time matching. Objective  $\sum_f (t_1 - t_2)^2$
- Let  $t_2' = a t_2 + b$ . Select  $a, b$  so as to minimize  $\sum_f (t_1 - t_2')^2$

# Geometric matching

- Make a graph. Peptide *a* in LCMS1 is linked to all peptides with identical  $m/z$ .
- Each edge has score proportional to  $t_1/t_2$
- Compute a maximum weight matching.
- The ratio of times of the matched pairs gives  $a$ .
- Rescale and compute the scaling factor



# Approach 2: Scan alignment

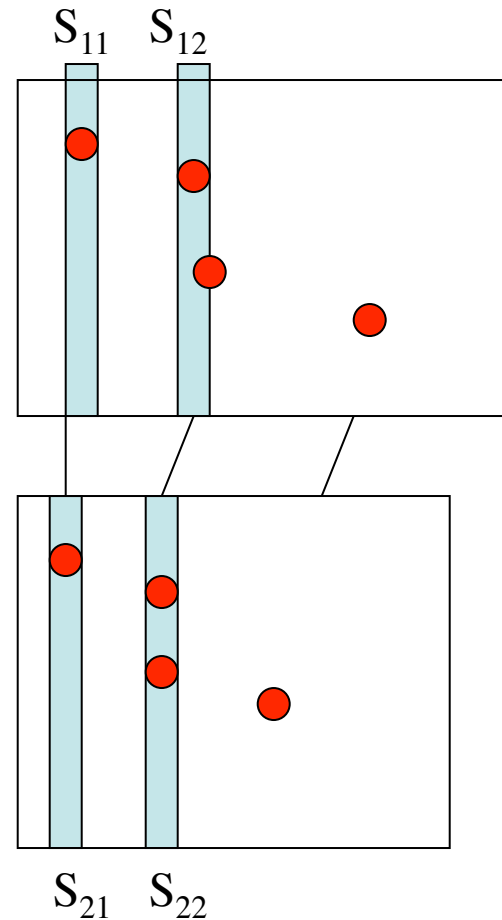
- Each time scan is a vector of intensities.
- Two scans in different runs can be scored for similarity (using a dot product)



$$S_{1i} = 10 \ 5 \ 0 \ 0 \ 7 \ 0 \ 0 \ 2 \ 9$$

$$S_{2j} = 9 \ 4 \ 2 \ 3 \ 7 \ 0 \ 6 \ 8 \ 3$$

$$M(S_{1i}, S_{2j}) = \sum_k S_{1i}(k) S_{2j}(k)$$

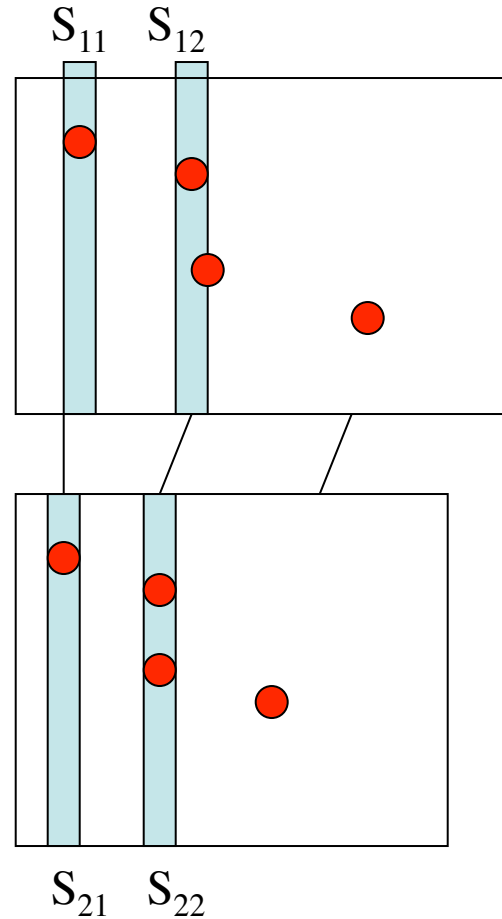


# Scan Alignment

- Compute an alignment of the two runs
- Let  $W(i,j)$  be the best scoring alignment of the first  $i$  scans in run 1, and first  $j$  scans in run 2

$$W(i,j) = \max \begin{cases} W(i-1,j-1) + M[S_{1i}, S_{2j}] \\ W(i-1,j) + \dots \\ W(i,j-1) + \dots \end{cases}$$

- Advantage: does not rely on feature detection.
- Disadvantage: Might not handle affine shifts in time scaling, but is better for local shifts



# Chemistry based methods for comparing peptides

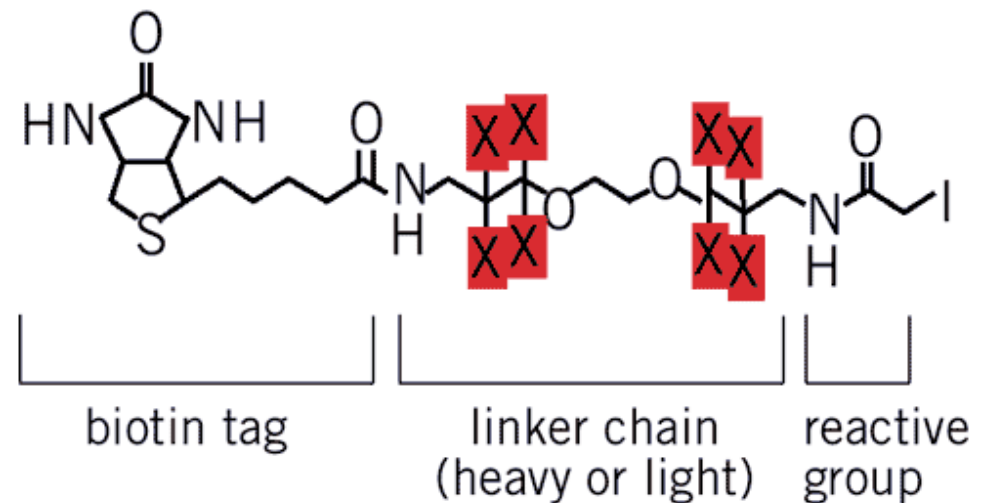
# ICAT

- The reactive group attaches to Cysteine
- Only Cys-peptides will get tagged
- The biotin at the other end is used to pull down peptides that contain this tag.
- The X is either Hydrogen, or Deuterium (Heavy)
  - Difference = 8Da

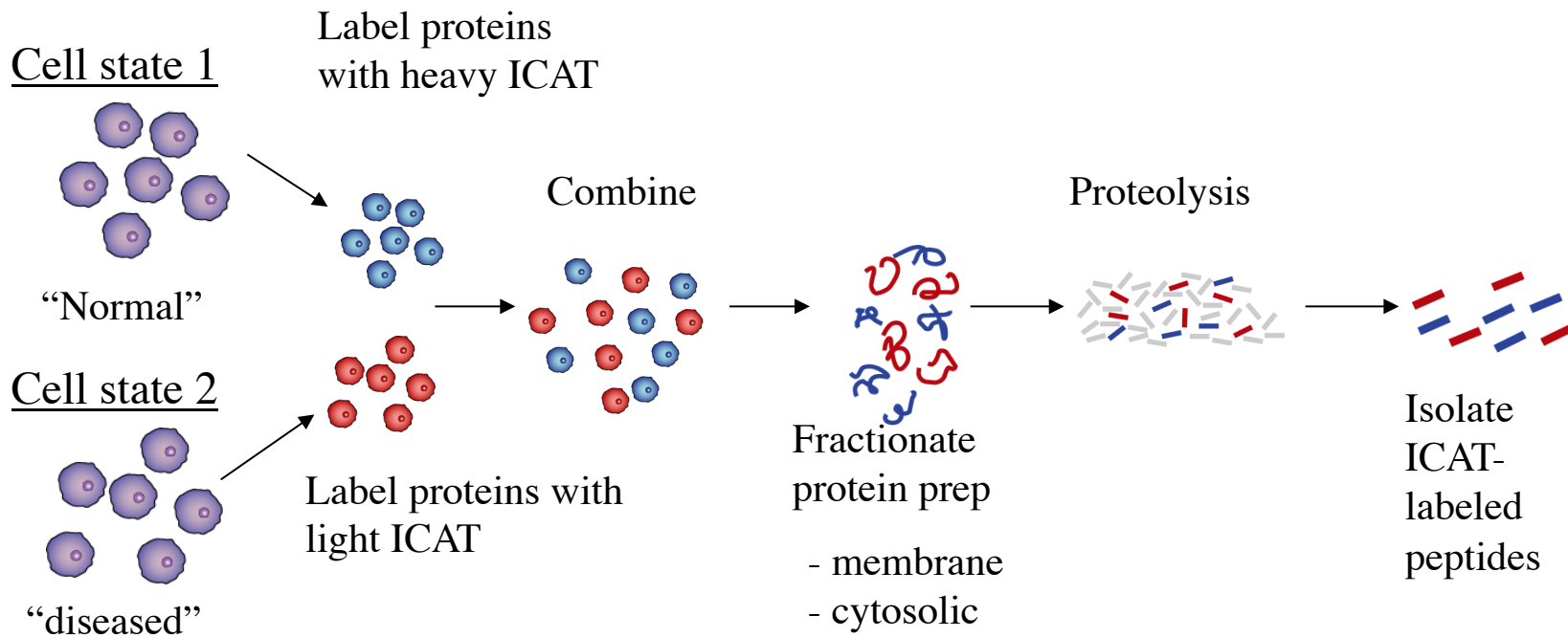
## Isotope-Coded Affinity Tags

heavy reagent: D8-ICAT Reagent (X=deuterium)

light reagent: D0-ICAT Reagent (X=hydrogen)



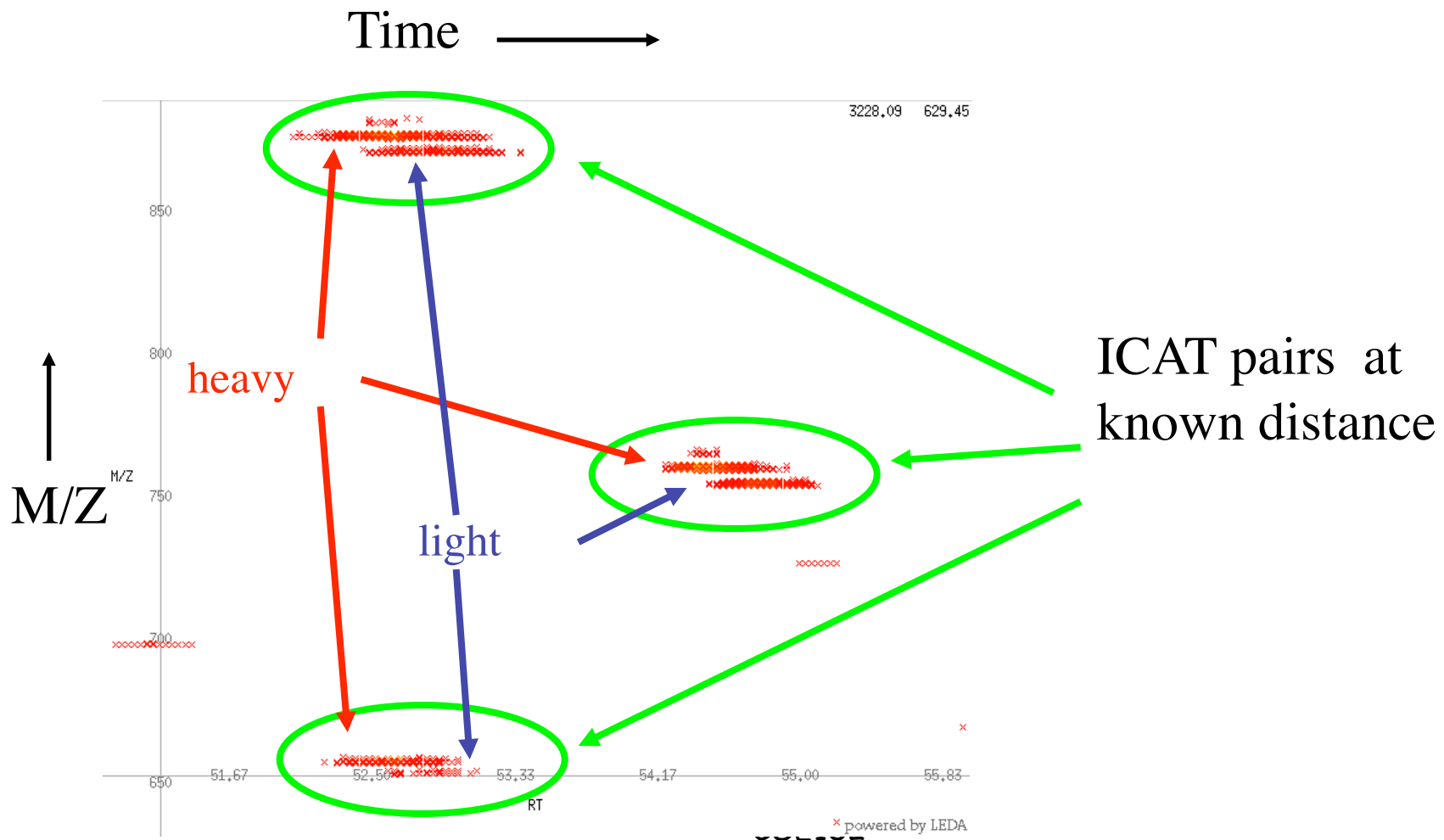
# ICAT



*Nat. Biotechnol.* 17: 994-999,1999

- ICAT reagent is attached to particular amino-acids (Cys)
- Affinity purification leads to simplification of complex mixture

# Differential analysis using ICAT

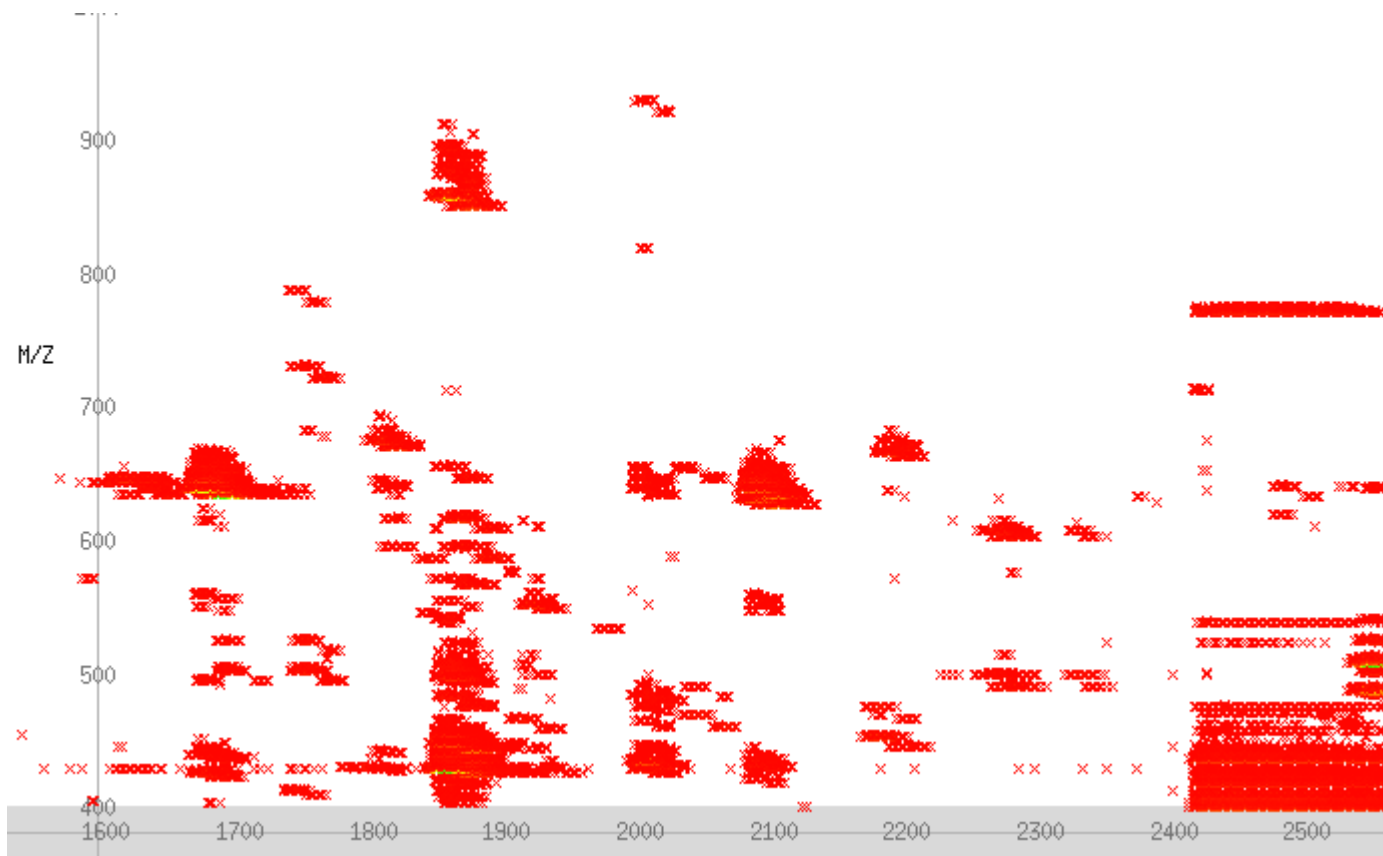


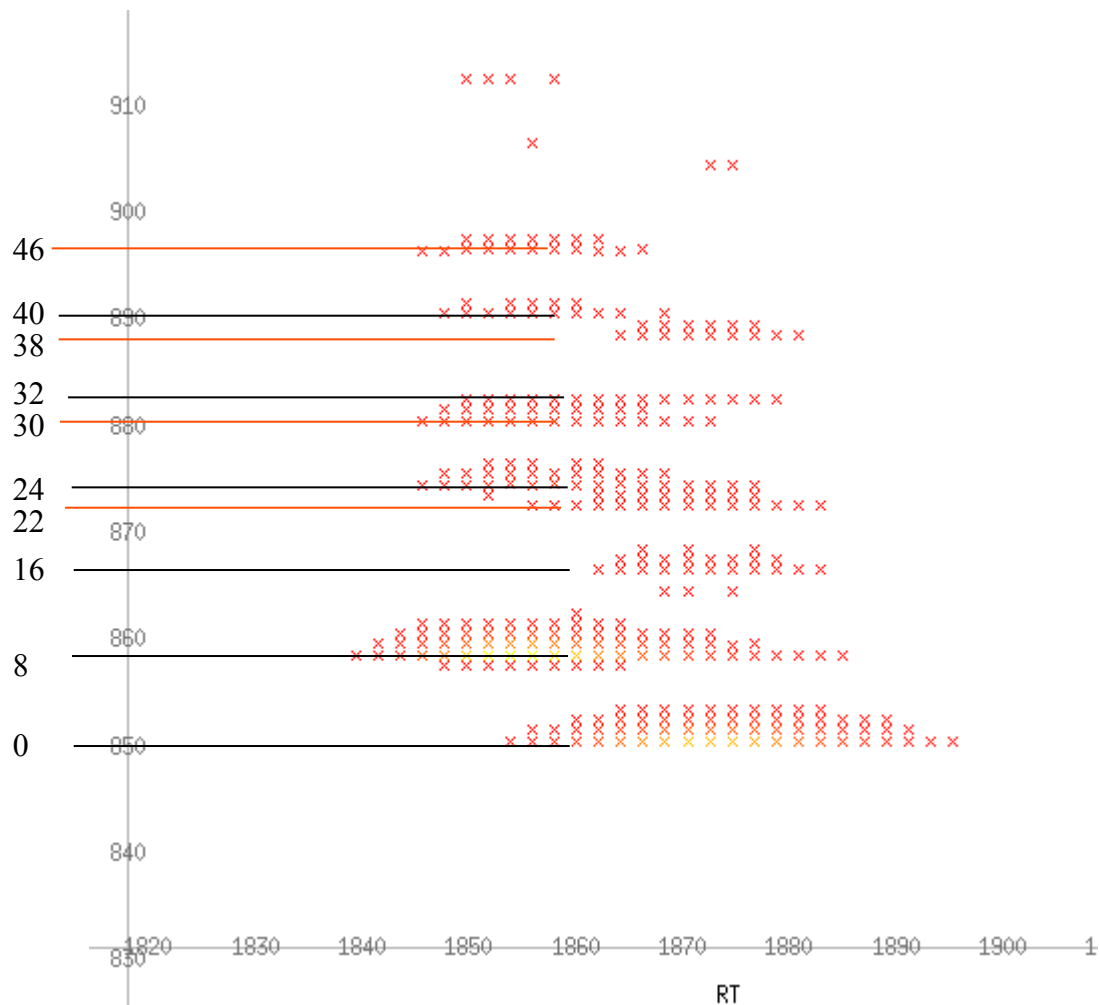
## ICAT issues

- The tag is heavy, and decreases the dynamic range of the measurements.
- The tag might break off
- Only Cysteine containing peptides are retrieved Non-specific binding to strepdavidin

# Serum ICAT data

MA13\_02011\_02\_ALL01Z3I9A\* Overview (exhibits 'stack-ups')





CSE182

# ICAT problems

- Tag is bulky, and can break off.
- Cys is low abundance
- $MS_2$  analysis to identify the peptide is harder.

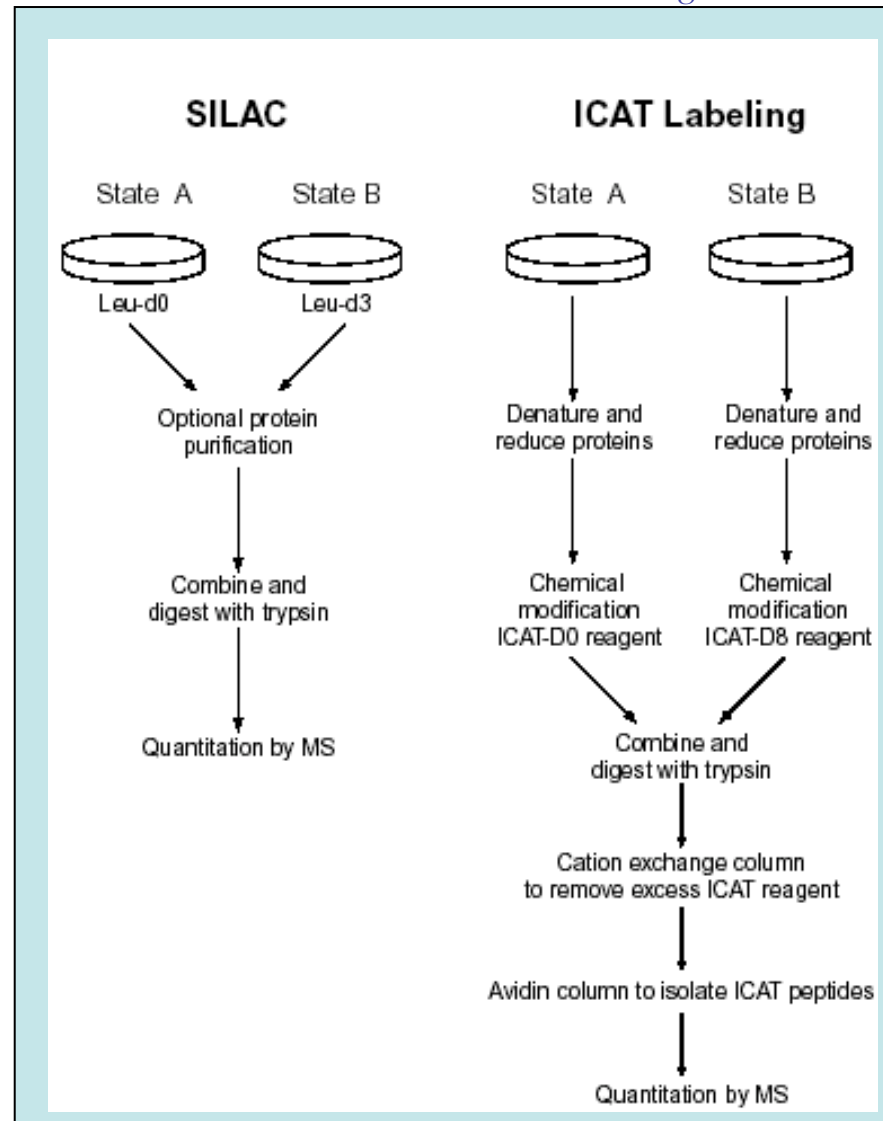
# SILAC

- A novel stable isotope labeling strategy
- Mammalian cell-lines do not 'manufacture' all amino-acids. Where do they come from?
- Labeled amino-acids are added to amino-acid deficient culture, and are incorporated into all proteins as they are synthesized
- No chemical labeling or affinity purification is performed.
- Leucine was used (10% abundance vs 2% for Cys)

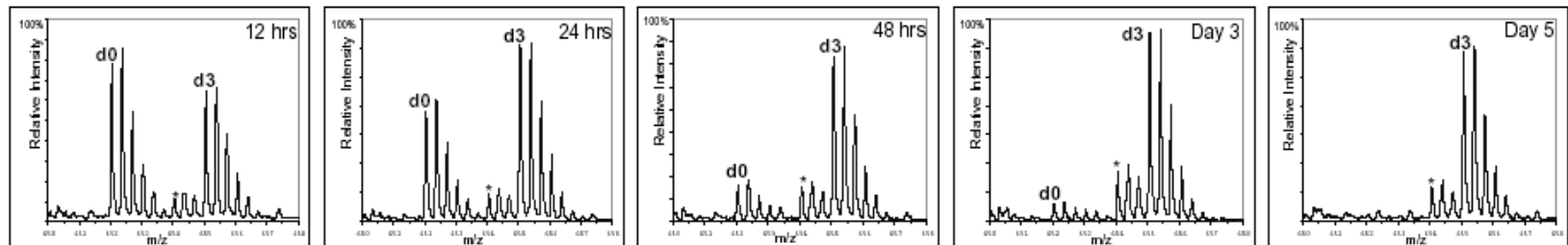
# SILAC vs ICAT

*Ong et al. MCP, 2002*

- Leucine is higher abundance than Cys
- No affinity tagging done
- Fragmentation patterns for the two peptides are identical
  - Identification is easier

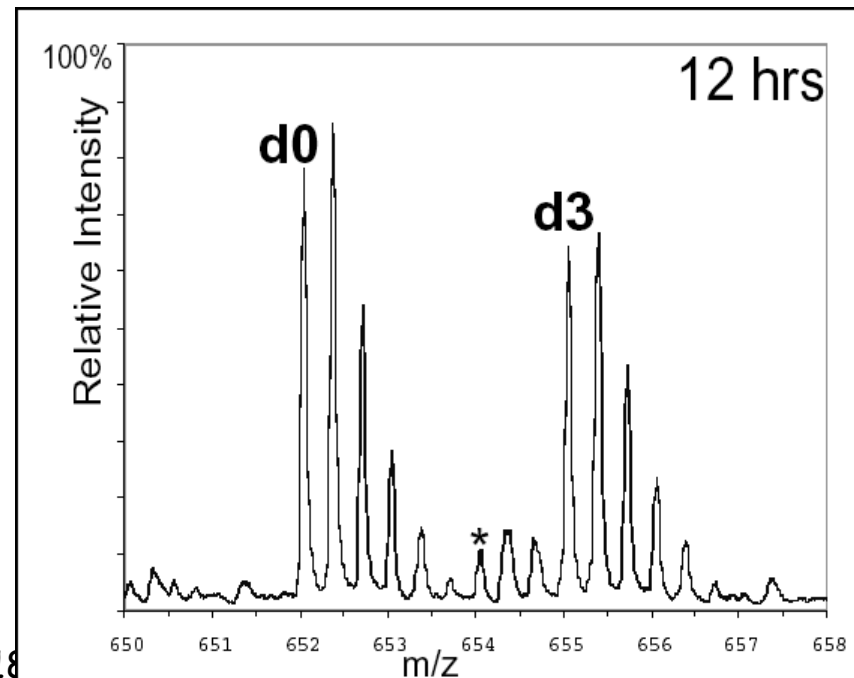


# Incorporation of Leu-d3 at various time points

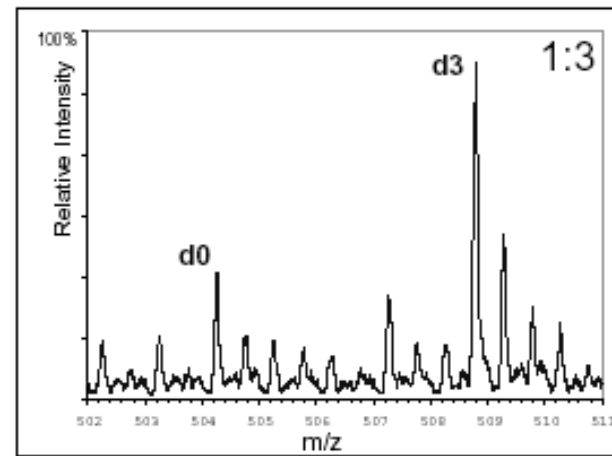
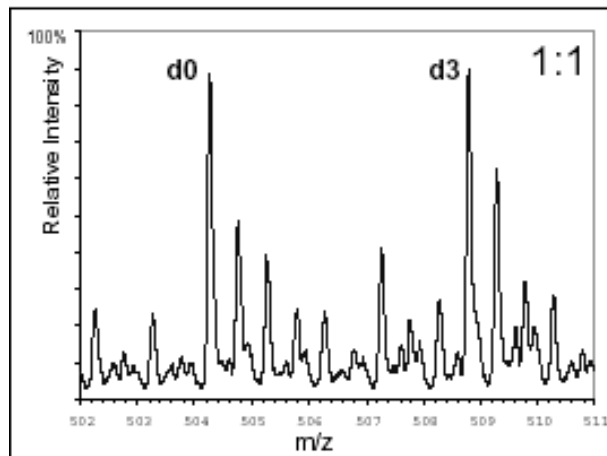


- Doubling time of the cells is 24 hrs.
- Peptide = VAPEEHPVLLTEAPLNPK
- What is the charge on the peptide?

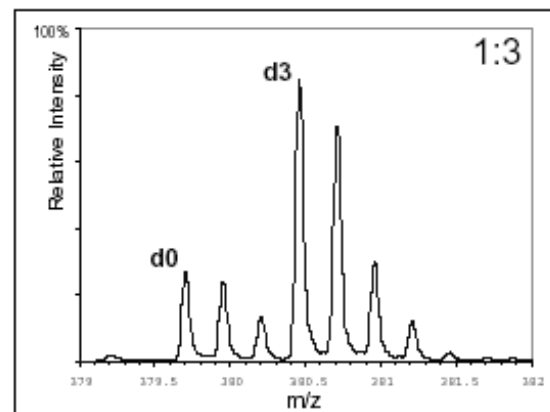
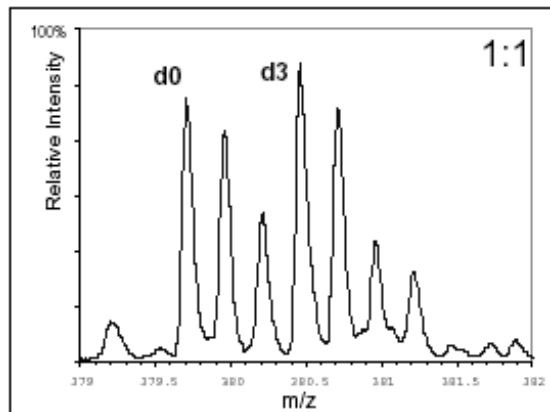
CSE18



# Quantitation on controlled mixtures



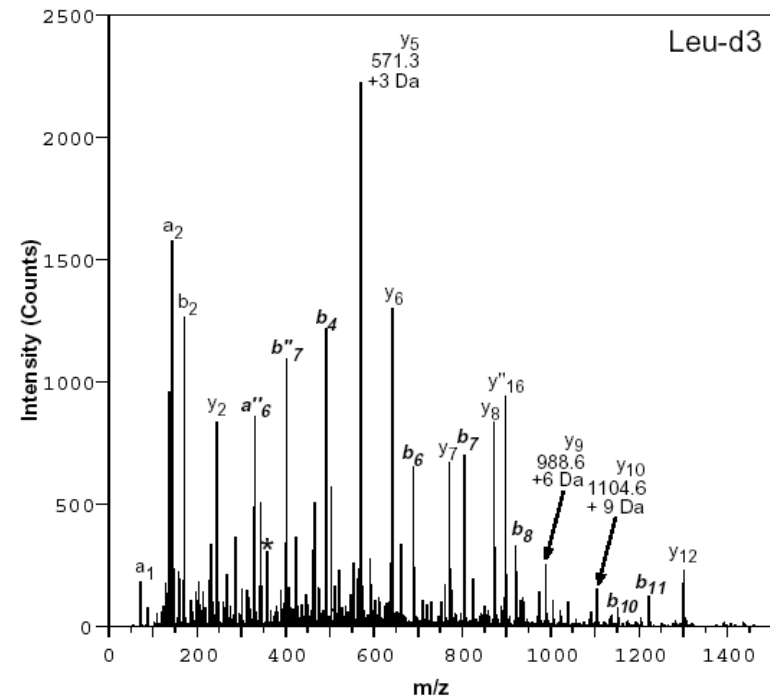
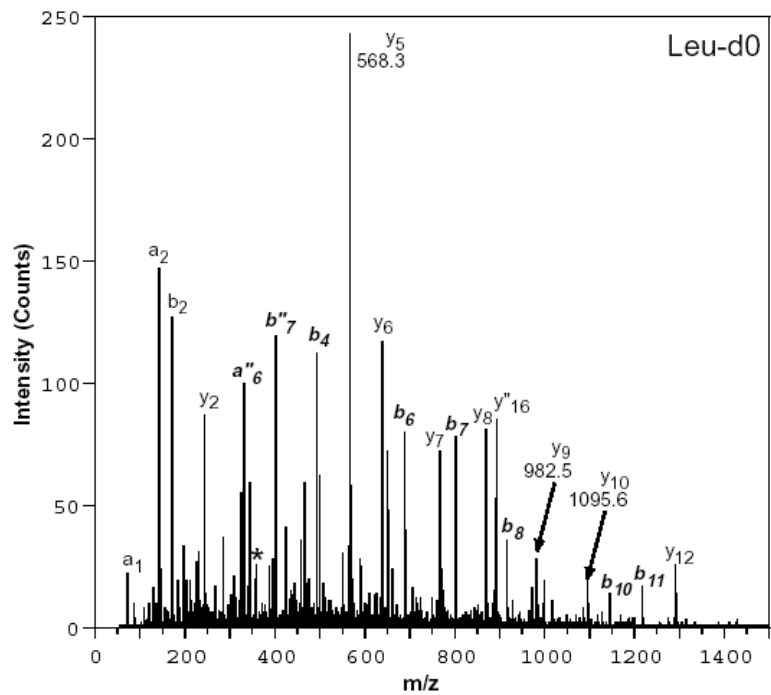
SCNCLLLK



IWHHTFYNELR

CSE182

# Identification



- MS/MS of differentially labeled peptides

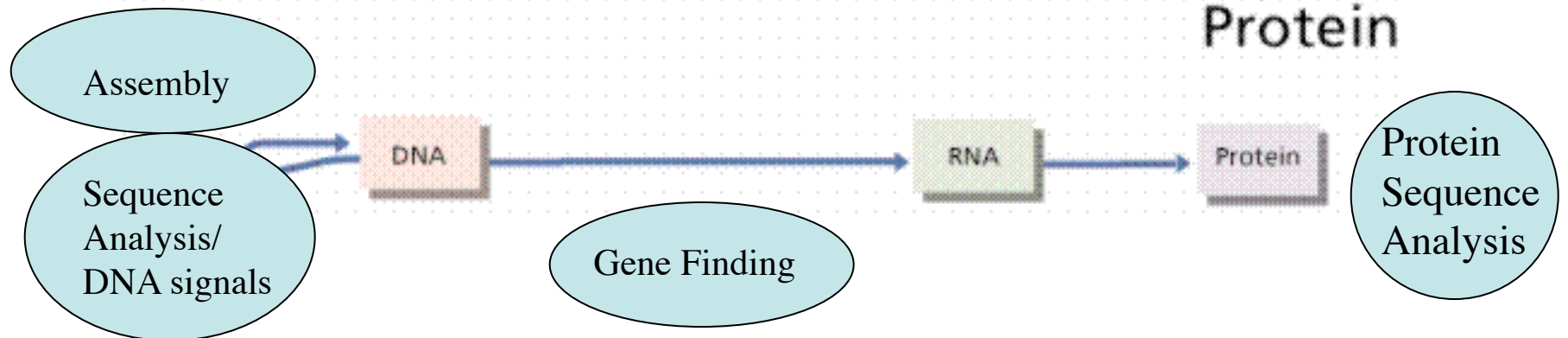
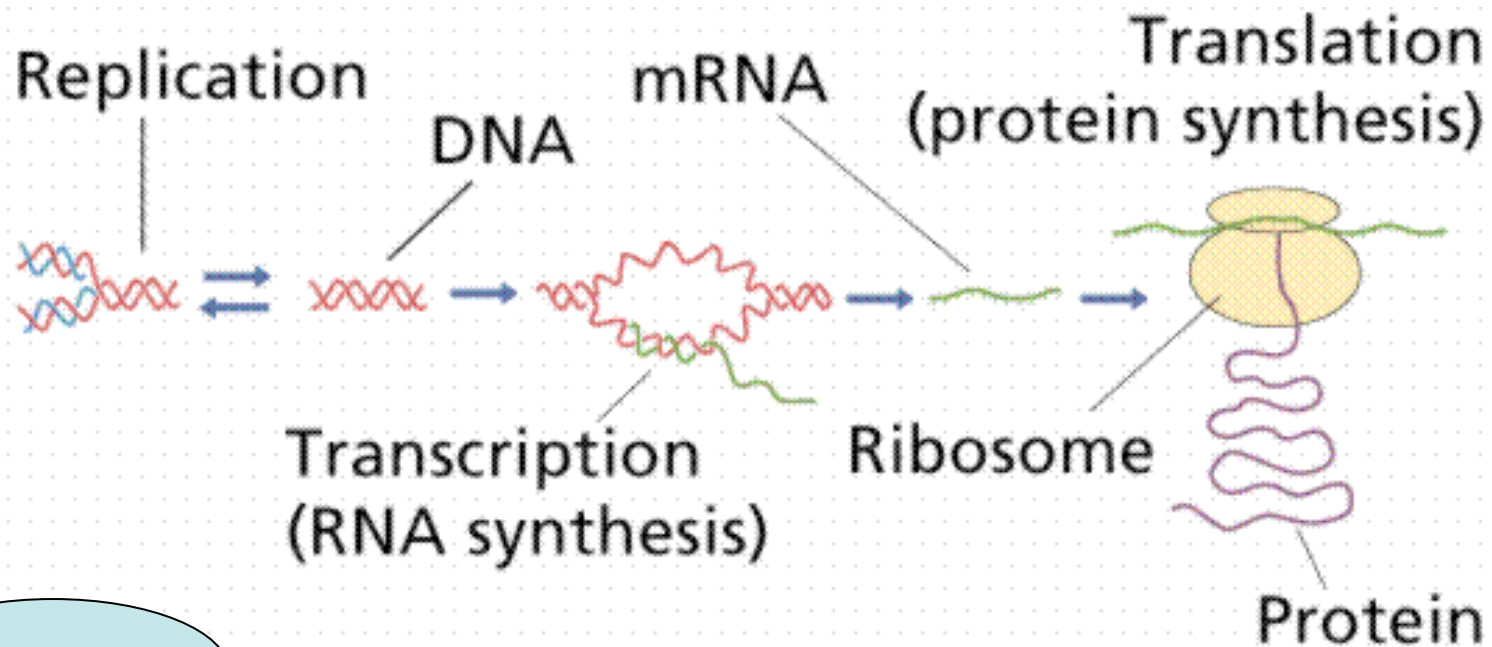
# Peptide Matching

- Computational: Under identical Liquid Chromatography conditions, peptides will elute in the same order in two experiments.
  - These peptides can be paired computationally
- SILAC/ICAT allow us to compare relative peptide abundances in a single run using an isotope tag.

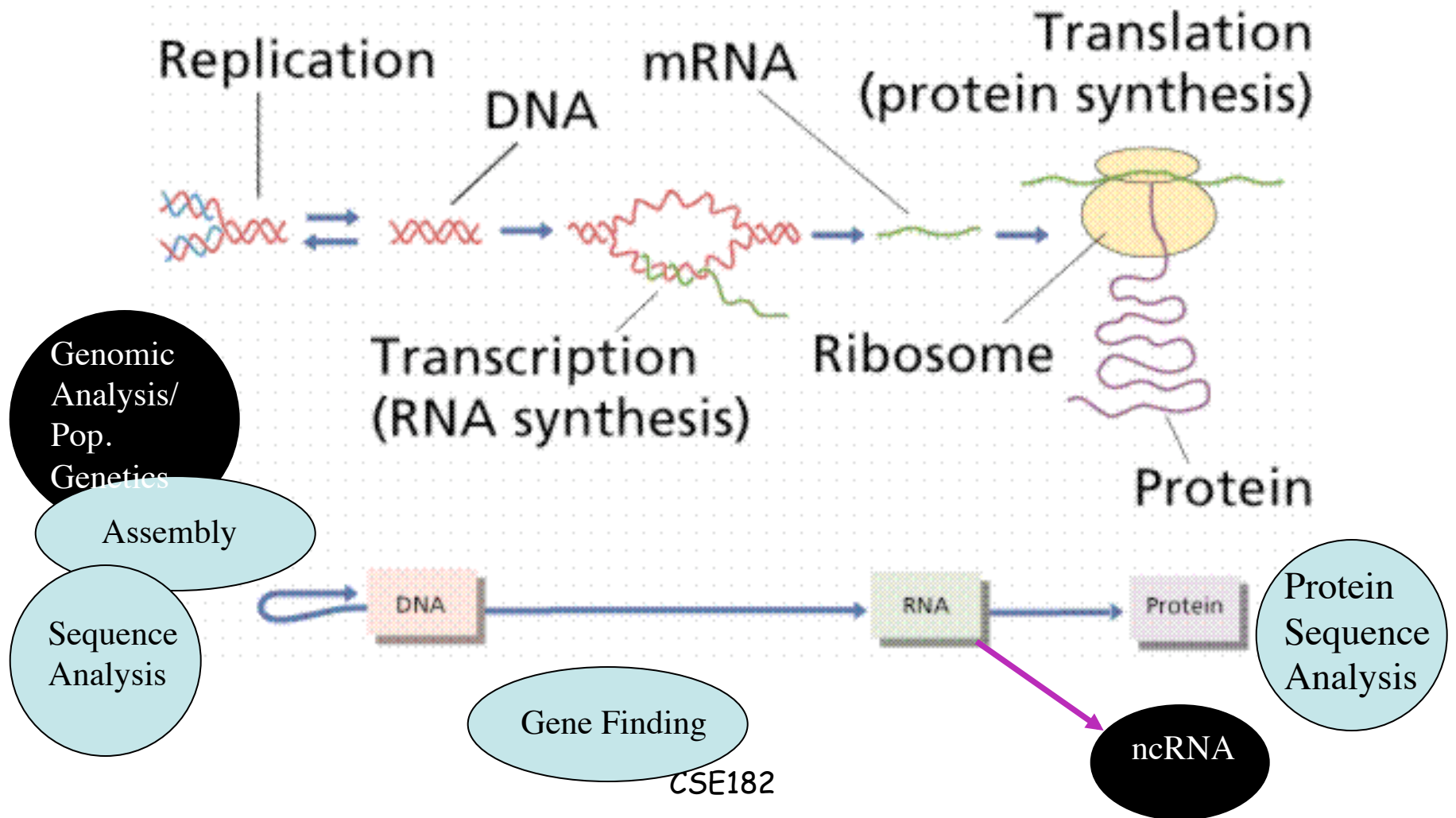
# MS quantitation Summary

- A peptide elutes over a mass range (isotopic peaks), and a time range.
- A 'feature' defines all of the peaks corresponding to a single peptide.
- Matching features is the critical step to comparing relative intensities of the same peptide in different samples.
- The matching can be done chemically (isotope tagging), or computationally (LCMS map comparison)

# Biol. Data analysis: Review

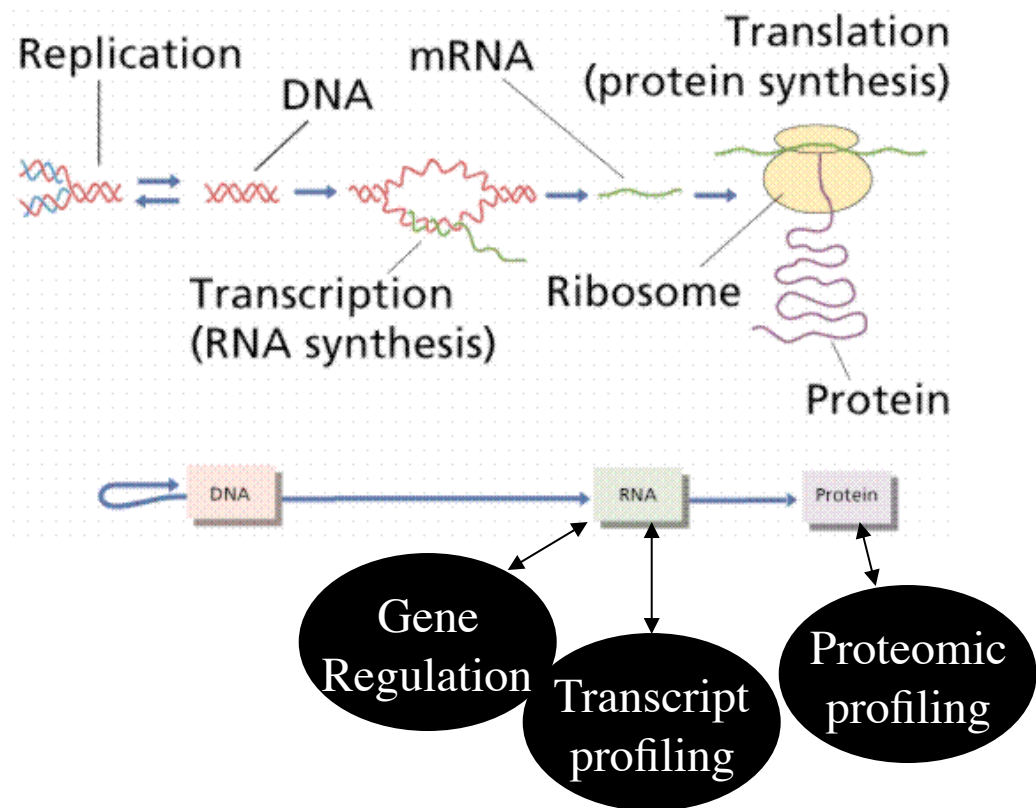


# Other static analysis is possible



## A Static picture of the cell is insufficient

- Each Cell is continuously active,
  - Genes are being transcribed into RNA
  - RNA is translated into proteins
  - Proteins are PT modified and transported
  - Proteins perform various cellular functions
- Can we probe the Cell dynamically?
  - Which transcripts are active?
  - Which proteins are active?
  - Which proteins interact?

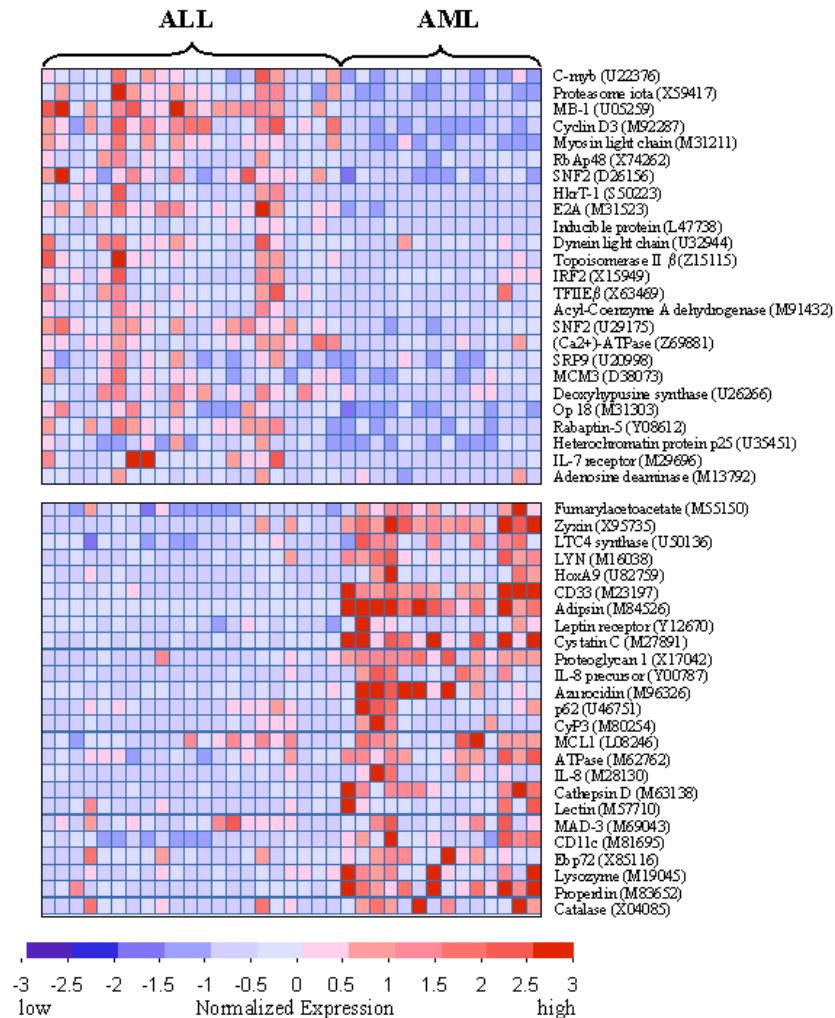


# Micro-array analysis

# The Biological Problem

- Two conditions that need to be differentiated, (Have different treatments).
  - EX: ALL (Acute Lymphocytic Leukemia) & AML (Acute Myelogenous Leukemia)
- Possibly, the set of expressed genes is different in the two conditions

## Independent Set



**Supplementary fig. 2. Expression levels of predictive genes in independent dataset. The expression levels of the 50 genes most highly correlated with the ALL-AML distinction in the initial dataset were determined in the independent dataset. Each row corresponds to a gene, with the columns corresponding to expression levels in different samples. The expression level of each gene in the independent dataset is shown relative to the mean of expression levels for that gene in the initial dataset. Expression levels greater than the mean are shaded in red, and those below the mean are shaded in blue. The scale indicates standard deviations above or below the mean. The top panel shows genes highly expressed in ALL, the bottom panel shows genes more highly expressed in AML.**

# Gene Expression Data

- Gene Expression data:
  - Each row corresponds to a gene
  - Each column corresponds to an expression value
- Can we separate the experiments into two or more classes?
- Given a training set of two classes, can we build a classifier that places a new experiment in one of the two classes.

	S <sub>1</sub>	S <sub>2</sub>				S
g						

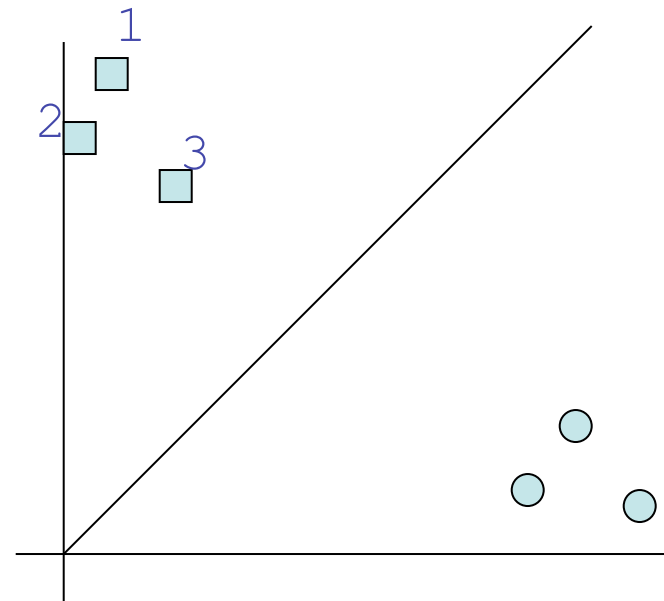
# Three types of analysis problems

- Cluster analysis/unsupervised learning
- Classification into known classes (Supervised)
- Identification of "marker" genes that characterize different tumor classes

# Supervised Classification: Basics

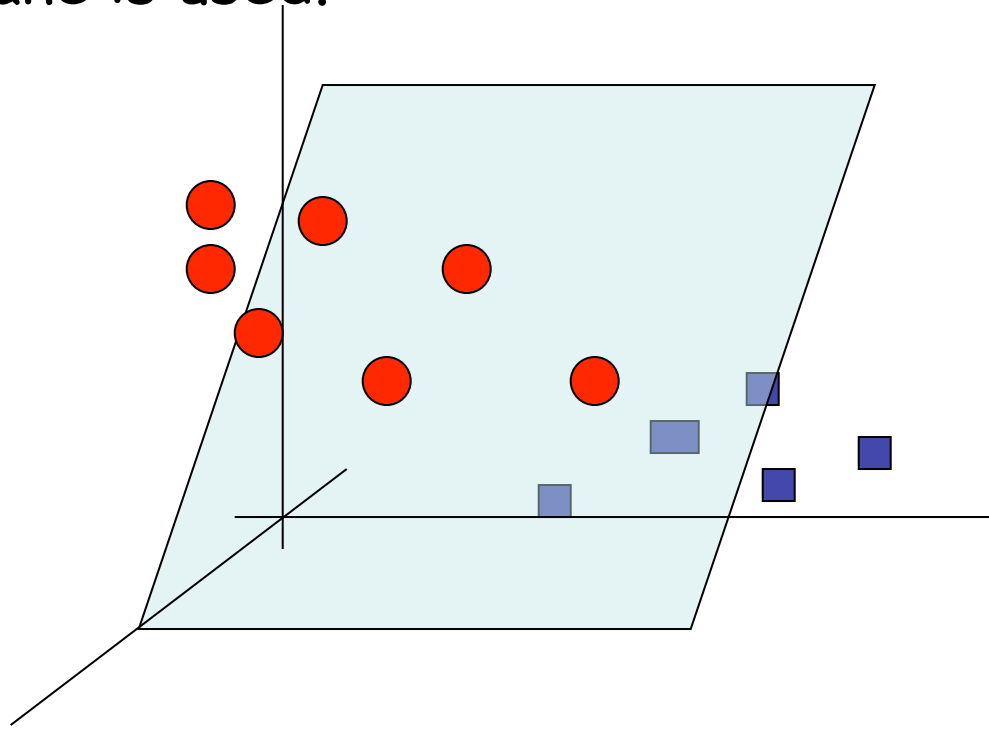
- Consider genes  $g_1$  and  $g_2$ 
  - $g_1$  is up-regulated in class A, and down-regulated in class B.
  - $g_2$  is up-regulated in class A, and down-regulated in class B.
- Intuitively,  $g_1$  and  $g_2$  are effective in classifying the two samples. The samples are linearly separable.

	1	2	3	4	5	6
$g_1$	1	.9	.8	.1	.2	.1
$g_2$	.1	0	.2	.8	.7	.9



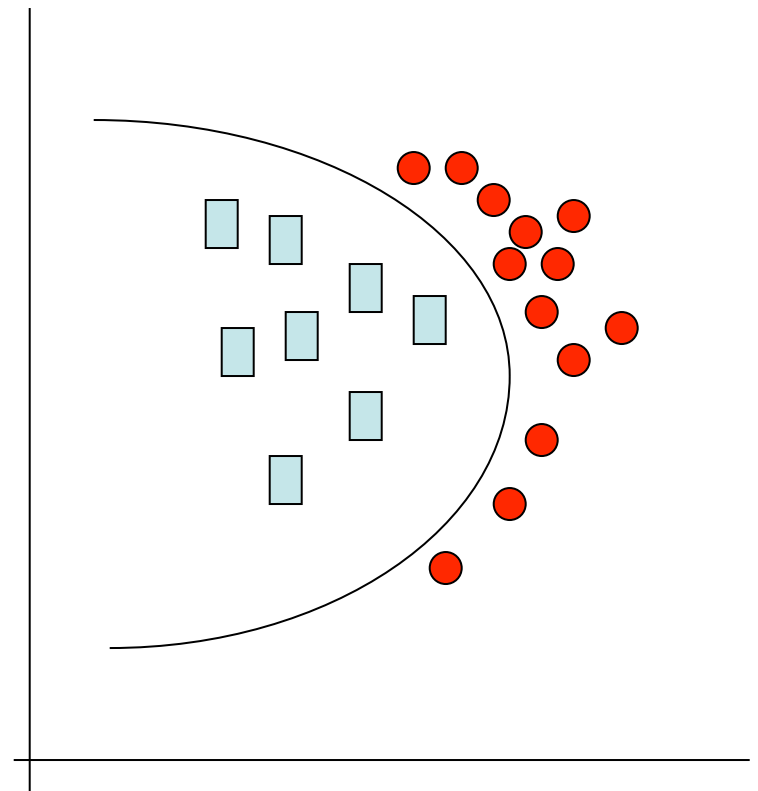
# Basics

- With 3 genes, a plane is used to separate (linearly separable samples). In higher dimensions, a hyperplane is used.



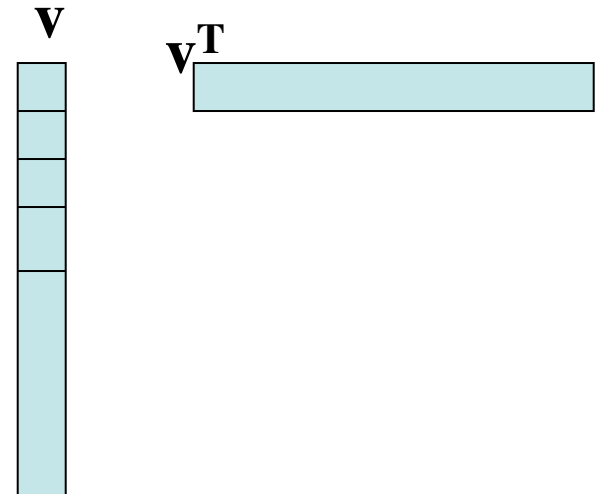
# Non-linear separability

- Sometimes, the data is not linearly separable, but can be separated by some other function
- In general, the linearly separable problem is computationally easier.



# Formalizing of the classification problem for micro-arrays

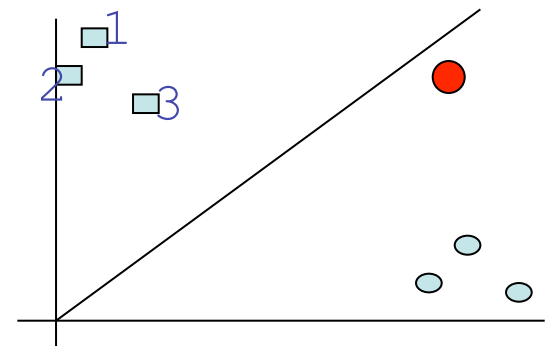
- Each experiment (sample) is a vector of expression values.
  - By default, all vectors  $\mathbf{v}$  are column vectors.
  - $\mathbf{v}^T$  is the transpose of a vector
- The genes are the dimension of a vector.
- Classification problem: Find a surface that will separate the classes



# Formalizing Classification

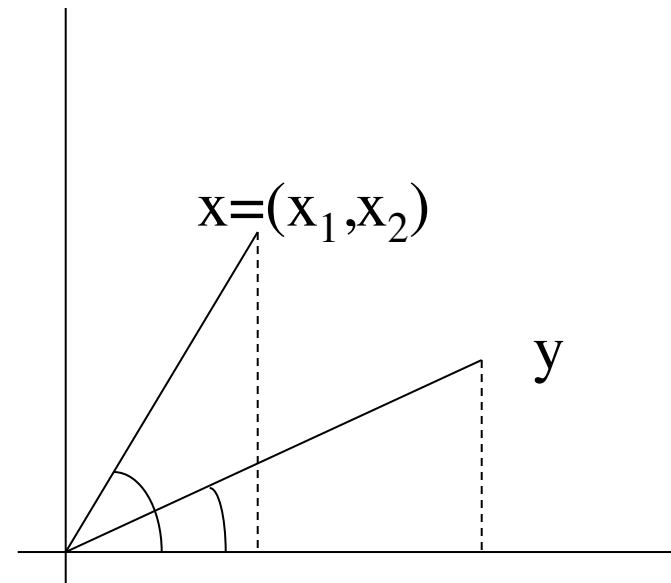
- Classification problem: Find a surface (hyperplane) that will separate the classes
- Given a new sample point, its class is then determined by which side of the surface it lies on.
- How do we find the hyperplane? How do we find the side that a point lies on?

	1	2	3	4	5	6
$g_1$	1	.9	.8	.1	.2	.1
$g_2$	.1	0	.2	.8	.7	.9



# Basic geometry

- What is  $\|\mathbf{x}\|_2$  ?
- What is  $\mathbf{x}/\|\mathbf{x}\|$
- Dot product?

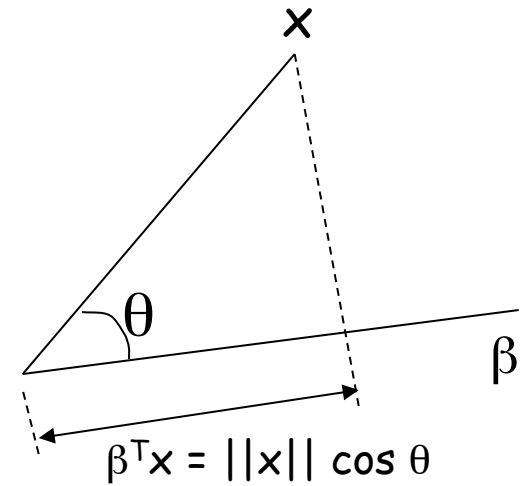


$$\begin{aligned}x^T y &= x_1 y_1 + x_2 y_2 \\ &= \|x\| \cdot \|y\| \cos \theta_x \cos \theta_y + \|x\| \cdot \|y\| \sin(\theta_x) \sin(\theta_y) \\ &= \|x\| \cdot \|y\| \cos(\theta_x - \theta_y)\end{aligned}$$

End of L14

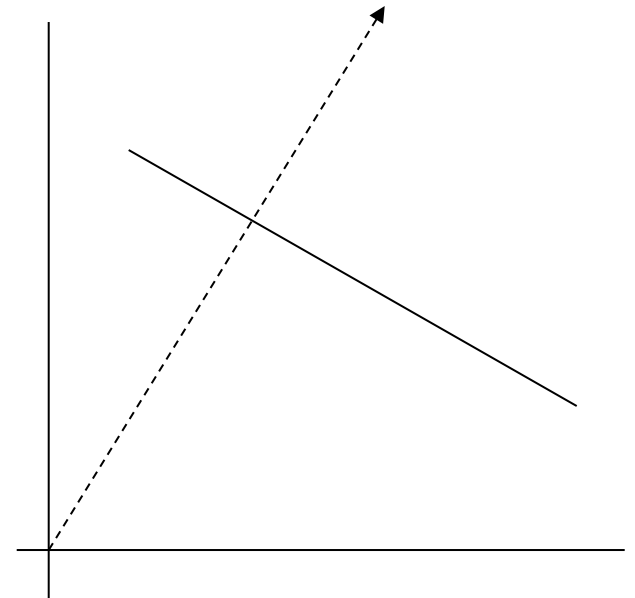
# Dot Product

- Let  $\beta$  be a unit vector.
  - $\|\beta\| = 1$
- Recall that
  - $\beta^T \mathbf{x} = \|\mathbf{x}\| \cos \theta$
- What is  $\beta^T \mathbf{x}$  if  $\mathbf{x}$  is orthogonal (perpendicular) to  $\beta$ ?



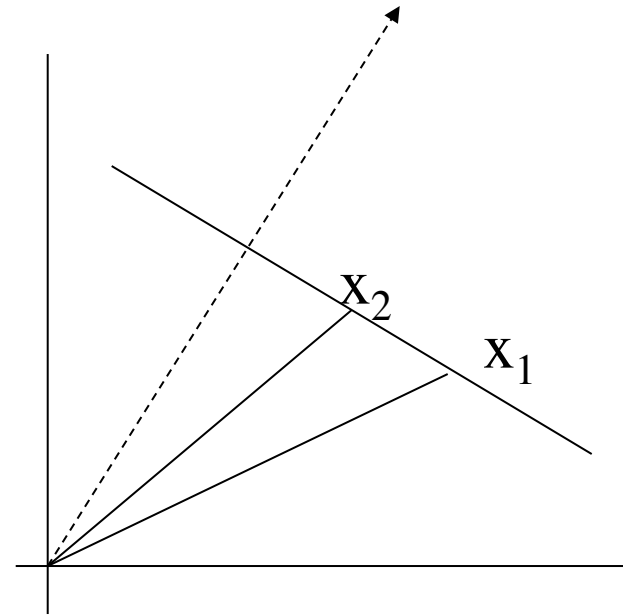
# Hyperplane

- How can we define a hyperplane  $L$ ?
- Find the unit vector that is perpendicular (normal to the hyperplane)



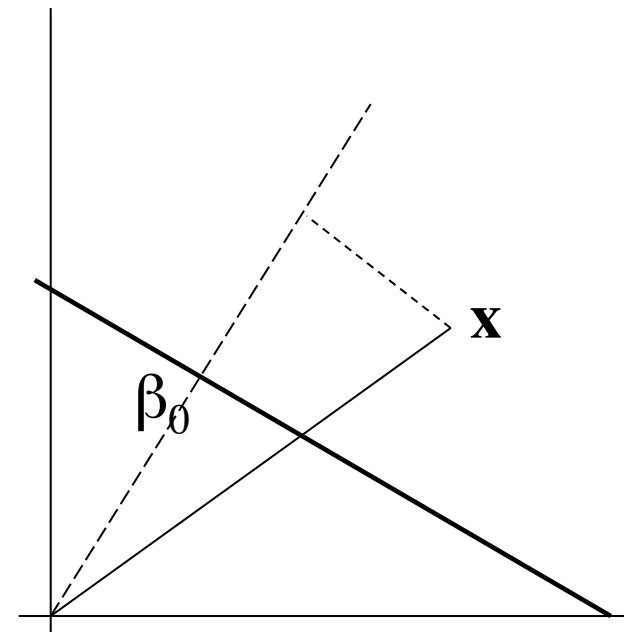
# Points on the hyperplane

- Consider a hyperplane  $L$  defined by unit vector  $\beta$ , and distance  $\beta_0$
- Notes;
  - For all  $x \in L$ ,  $x^T \beta$  must be the same,  $x^T \beta = \beta_0$
  - For any two points  $x_1, x_2$ ,
    - $(x_1 - x_2)^T \beta = 0$



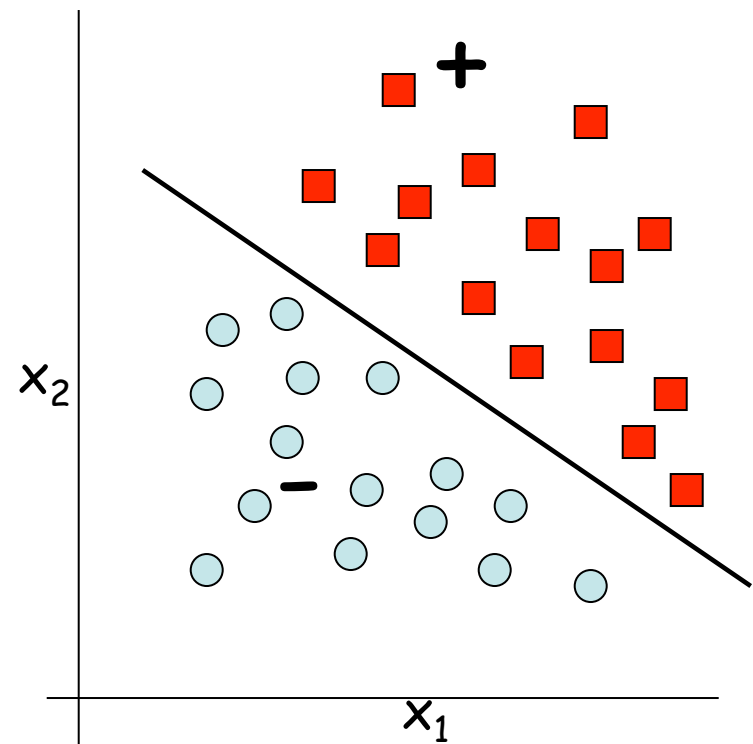
# Hyperplane properties

- Given an arbitrary point  $x$ , what is the distance from  $x$  to the plane  $L$ ?
  - $D(x,L) = (\beta^T x - \beta_0)$
- When are points  $x_1$  and  $x_2$  on different sides of the hyperplane?



# Separating by a hyperplane

- Input: A training set of +ve & -ve examples
- Goal: Find a hyperplane that separates the two classes.
- Classification: A new point  $x$  is +ve if it lies on the +ve side of the hyperplane, -ve otherwise.
- The hyperplane is represented by the line
- $\{x: -\beta_0 + \beta_1 x_1 + \beta_2 x_2 = 0\}$

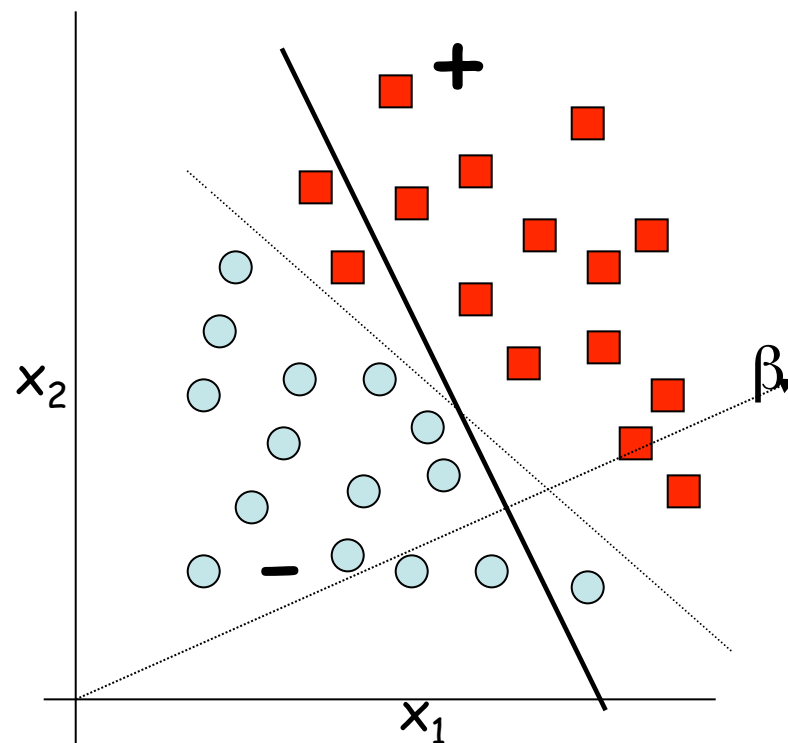


# Error in classification

- An arbitrarily chosen hyperplane might not separate the test. We need to minimize a mis-classification error
- Error: sum of distances of the misclassified points.
- Let  $y_i = -1$  for +ve example  $i$ ,
  - $y_i = 1$  otherwise.

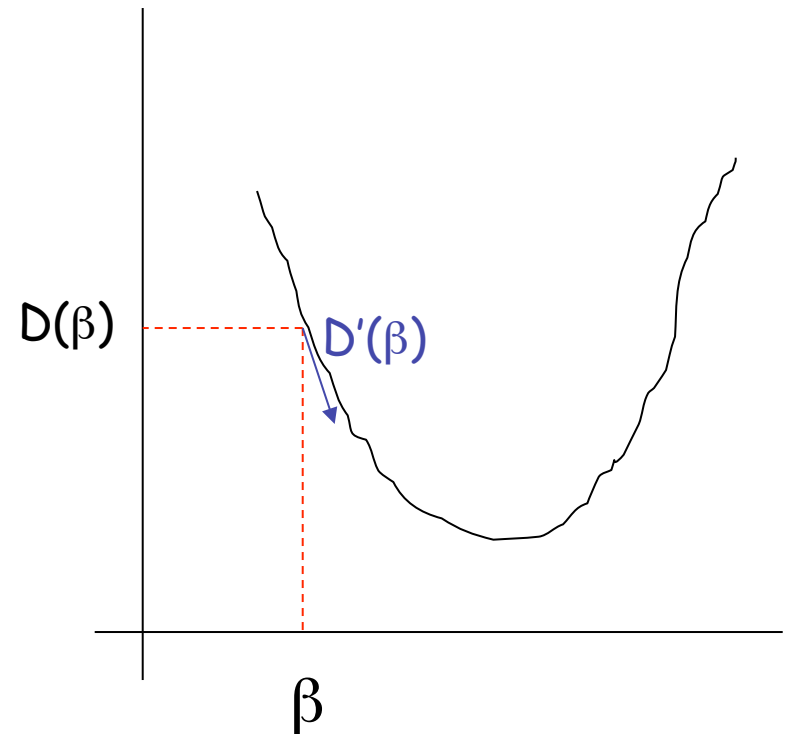
$$D(\beta, \beta_0) = \sum_{i \in M} y_i (x_i^T \beta + \beta_0)$$

- Other definitions are also possible.



# Gradient Descent

- The function  $D(\beta)$  defines the error.
- We follow an iterative refinement. In each step, refine  $\beta$  so the error is reduced.
- Gradient descent is an approach to such iterative refinement.



$$\beta \leftarrow \beta - \rho \cdot D'(\beta)$$

# Rosenblatt's perceptron learning algorithm

$$D(\beta, \beta_0) = \sum_{i \in M} y_i (x_i^T \beta + \beta_0)$$

$$\frac{\partial D(\beta, \beta_0)}{\partial \beta} = \sum_{i \in M} y_i x_i$$

$$\frac{\partial D(\beta, \beta_0)}{\partial \beta_0} = \sum_{i \in M} y_i$$

$$\Rightarrow \text{Update rule : } \begin{pmatrix} \beta \\ \beta_0 \end{pmatrix} = \begin{pmatrix} \beta \\ \beta_0 \end{pmatrix} - \rho \begin{pmatrix} \sum_{i \in M} y_i x_i \\ \sum_{i \in M} y_i \end{pmatrix}$$

# Classification based on perceptron learning

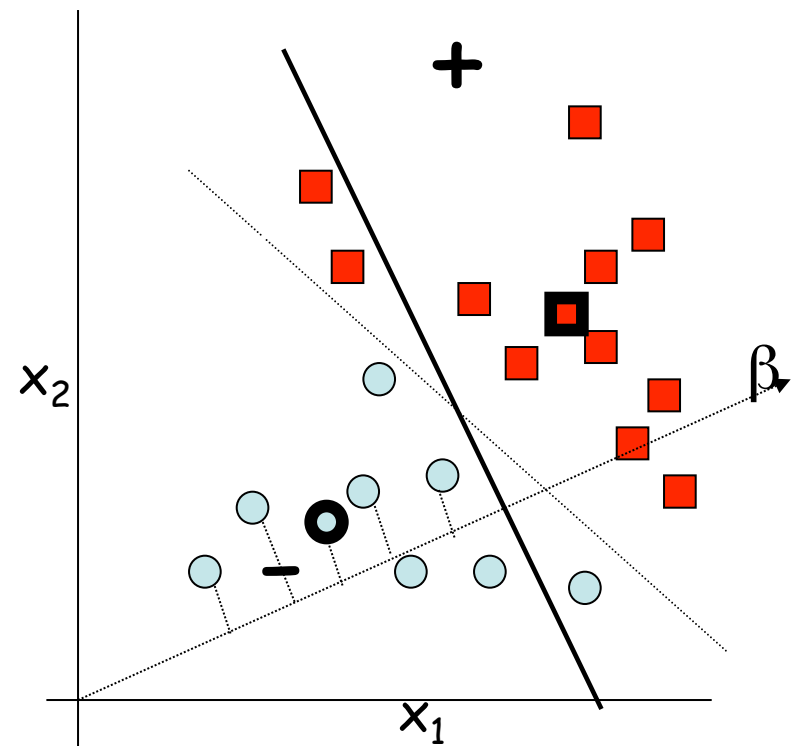
- Use Rosenblatt's algorithm to compute the hyperplane  $L=(\beta,\beta_0)$ .
- Assign  $x$  to class 1 if  $f(x) \geq 0$ , and to class 2 otherwise.

# Perceptron learning

- If many solutions are possible, it does not choose between solutions
- If data is not linearly separable, it does not terminate, and it is hard to detect.
- Time of convergence is not well understood

# Linear Discriminant analysis

- Provides an alternative approach to classification with a linear function.
- Project all points, including the means, onto vector  $\beta$ .
- We want to choose  $\beta$  such that
  - Difference of projected means is large.
  - Variance within group is small



## LDA Cont'd

$$\tilde{m}_1 = \frac{1}{n_1} \sum_x \beta^T x = w^T m_1$$

$$\text{Scatter between samples: } |\tilde{m}_1 - \tilde{m}_2|^2 = |\beta^T (m_1 - m_2)|^2$$

$$|\tilde{m}_1 - \tilde{m}_2|^2 = \beta^T S_B \beta$$

$$\text{scatter within sample: } \tilde{s}_1^2 + \tilde{s}_2^2$$

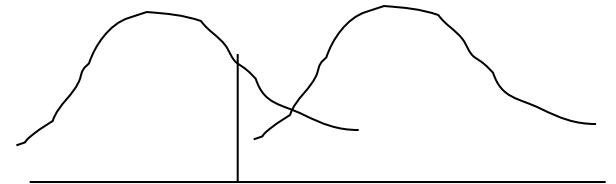
$$\text{where, } \tilde{s}_1^2 = \sum_y (y - \tilde{m}_1)^2 = \sum_{x \in D_1} (\beta^T (x - m_1))^2 = \beta^T S_1 \beta$$

$$\tilde{s}_1^2 + \tilde{s}_2^2 = \beta^T (S_1 + S_2) \beta = \beta^T S_w \beta$$

$$\text{Fisher Criterion } \max_{\beta} \frac{\beta^T S_B \beta}{\beta^T S_w \beta}$$

# Maximum Likelihood discrimination

- Suppose we knew the distribution of points in each class.
  - We can compute  $\Pr(x|\omega_i)$  for all classes  $i$ , and take the maximum



# ML discrimination

- Suppose all the points were in 1 dimension, and all classes were normally distributed.

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$$

$$\Pr(\omega_i | x) = \frac{\Pr(x | \omega_i) \Pr(\omega_i)}{\sum_j \Pr(x | \omega_j) \Pr(\omega_j)}$$

$$\begin{aligned} g_i(x) &= \ln(\Pr(x | \omega_i)) + \ln(\Pr(\omega_i)) \\ &\cong \frac{-(x - \mu_i)^2}{2\sigma_i^2} + \ln(\Pr(\omega_i)) \end{aligned}$$

## ML discrimination recipe

- We know the distribution for each class, but not the parameters
- Estimate the mean and variance for each class.
- For a new point  $x$ , compute the discrimination function  $g_i(x)$  for each class  $i$ .
- Choose  $\operatorname{argmax}_i g_i(x)$  as the class for  $x$